United States Department of Agriculture

Animal and Plant Health Inspection Service

DWRC Research Report No. 11-55-004

Interpolation in the Nonparametric Density Estimator of Patil et al.

By R. M. Engeman, L. F. Pank, R. T. Sugihara, and W. E. Dusenberry

Drs. Engeman and Dusenberry are with the Denver Wildlife Research Center, a unit of the U.S. Department of Agriculture's Animal and Plant Health Inspection Service's Animal Damage Control program. Mr. Sugihara is with the Center at its Hilo, HI, field station. Mr. Pank works for the U.S. Department of the Interior's Alaska Fish and Wildlife Research Center in Anchorage.

Introduction

Plotless (distance) methods have received considerable attention during the past 20 years for the estimation of density for a population of stationary objects. Such plotless density estimators (PDE's) have usually been developed under the assumption that the population of objects is randomly (Poisson) distributed throughout the area to be measured. However, objects in nature seldom follow a random spatial distribution, and estimators based on this assumption usually do not perform well when the assumption does not hold (e.g., Diggle 1975).

Patil et al. (1979) developed a nonparametric PDE, based on fundamental theoretical results, to overcome the lack of robustness of most PDE's over different spatial patterns. In this report, we consider a modification to the nonparametric PDE given by Patil et al. (henceforth referred to as the PBK estimator).

To apply the PBK estimator, n points are randomly sampled in the area of interest. The distance, R, to the closest individual of interest is measured from each point. For each sample point, the search area to locate the closest individual is calculated as $U = \pi R^2$. The n search areas are ordered as $U_{(i)}$ where (i) indicates the i^{th} order statistic. The PBK estimator, as indicated by Patil et al. (1979), is a special case of an estimation procedure developed by Loftsgaarden and Quesenberry (1965) and is written as

$$\hat{f}(0) = (k(n)/n)/U_{([k(n)])} \qquad (1)$$

where $\hat{f}(0)$ is the estimated density, [ ] indicates the greatest integer function, and k(n) is a sequence of real numbers such that

$$\lim_{n\to\infty} k(n) = \infty \quad \text{and} \quad \lim_{n\to\infty} k(n)/n = 0. \qquad (2)$$

Patil et al. (1979) recommended that the form of the sequence be defined as

$$k(n) = n^p, \text{ where } p < 1. \qquad (3)$$

They specifically encouraged the use of an estimator based on the above sequence with p = 0.5. In their more recent article (Patil et al. 1982), they indicated that p = 2/3 is asymptotically optimal.

A modification to the PBK estimator is apparent from equation (1). Rather than selecting the $[k(n)]^{th}$ order statistic from among the ordered U's, we consider calculating the $\hat{f}(0)$ in equation (1) by linear interpolation between the ordered $U_{(i)}$ and $U_{(i+1)}$ where $i \le k(n) \le i + 1$. Intuitively, the application of interpolation to the estimation procedure would be of greatest benefit with small sample sizes, where the magnitude of the differences between consecutive U's tends to be larger.

Simulation Study

To examine the effect of interpolation in the small sample size situation, we performed a Monte Carlo simulation study with six populations of objects, each with a different spatial pattern. We identify the six spatial patterns for the simulated populations as random, triangular, regular, aggregate, double clumped, and field. The *random* pattern (also called Poisson in recognition that the points are distributed as a two-dimensional Poisson process) was simulated by generating random x-y coordinates. The *triangular* pattern was generated so that the population members were located at the vertices of a lattice of equilateral triangles. Because each object is equidistant from its six neighbors, this distribution is also called a hexagonal distribution. The *regular* spatial pattern was generated by dividing the area into a grid of rectangles, the same number as individuals to be in the population. The population members were then situated by randomly locating one individual in each rectangle. For the aggregate pattern, the centers of clumps were randomly located. In addition to the clump center point, "offspring" for the clumps were located about the center (parent) point, using coordinates randomly generated from the standard bivariate normal distribution. This procedure tends to concentrate the members of the clump close to the center point. The aggregate patterns approximate many naturally occurring biological population patterns, including rat damage patterns in sugarcane. The pattern we label as *double clumped* is a second order aggregation that was generated in a similar fashion to the aggregate pattern. The difference is that for the double clumped pattern, the individuals in the clumps of the aggregate pattern are used for center points (parents) for subclumps of two individuals. The two individuals of the subclumps include the parent plus one other point (offspring) randomly generated from the standard bivariate normal distribution. The *field* population was provided by our most immediate application for PDE's, estimating the density of rat-damaged internodes on cane stalks in sugarcane fields. This spatial pattern was produced by locating and recording each damaged internode in a section of a cane field on the island of Hawaii.

In this study, we did not consider truncation formulae for restricted search areas. We also avoided edge effects by rejecting sample points where the search area encountered an edge prior to finding an object of interest. The true density for each simulated distribution was approximately 20, which approximates densities from actual field data. The sample size used in the simulations was also motivated from in-field experience. Closest individual measurements for density estimation are most efficient, relative to plot sampling, at smaller sample sizes (Holgate 1964). Also, cane fields are very arduous to sample, and personnel to do the sampling are usually limited. Therefore, about 10 samples per field are all that are taken. Thus, after the populations were generated, samples from 10 randomly placed points were selected from each population. The distance from each random point to the closest individual was measured, and the corresponding search areas were calculated. From each data set of 10 distances, a density estimate was calculated.

Results

To investigate the effects that interpolation and the value of p have on the estimation procedure, we considered three values of p: 0.5, 0.6, and 0.7. For each value of p, an interpolated and an uninterpolated estimator were considered, resulting in six estimators (three pairs) for simulation.

We considered the use of real-world data in the simulation to be vitally important if we were to potentially give managers of sugarcane fields a sampling tool with which they could make management decisions. However, incorporating the field data in the simulations posed a minor dilemma as to what the optimal simulation methods should be. We would have preferred to generate a new population for each of the spatial patterns at each iteration of the simulation. However, having only one set of field data available precluded this. Also, the triangular pattern (with no randomness among the point locations) would look the same each iteration, and the square pattern would have only minor changes relative to the overall pattern. Based on these considerations, our choice for treating the six patterns in a consistent manner was to generate each population once and then use a new set of sampling points at each iteration. The sampling procedure and density estimation were repeated 300 times for each pattern. The simulations were conducted in FORTRAN on a [illegible]C 382/S with an MK-II microfile CPU and disk drive.

We compared the simulation results for the estimators by examining their relative root mean square error (RRMSE). If we let D equal the actual density and $\hat{D} = \hat{f}(0)$, then the (observed) RRMSE is calculated as

$$RRMSE = [(\Sigma(\hat{D}-D)^2/I)D^2]^{1/2} \quad (4)$$

where I is the number of replications (iterations) in the simulation.

The RRMSE results and the true densities from the simulation are given in table 1. These results indicate that interpolation consistently enhances the estimation procedure. Across the six populations, the interpolated RRMSE is smaller (generally much smaller) for each value of p than the uninterpolated value in all but two instances, and in one of these two instances the interpolated RRMSE is only 0.007 larger. However, we cannot explain the results for the other instance, 0.6I for the random pattern, where the original version outperforms the interpolated modification. Over all values, the RRMSE is 23 percent smaller for the interpolated estimators.

Based on equations (1) and (3), one would expect interpolation to have the greatest effect when n^p is slightly smaller than an integer. If it is slightly larger than an integer, there would be little effect. For example, $10^{0.7} = 5.01$. Using this value for interpolation would not result in much improvement over the greatest integer function results with $[10^{0.7}] = 5$. However, for $10^{0.5}$ and $10^{0.6}$, interpolation would be expected to have a more dramatic effect because $10^{0.5} = 3.16$ but $[10^{0.5}] = 3$ and $10^{0.6} = 3.98$ but $[10^{0.6}] = 3$. These expectations generally hold true in the results given in table 1.

Table 1—RRMSE's for the modifications to the PBK estimator for sample size n=10 and 300 iterations in the simulation

Spatial pattern	True density	Estimator[1]					
		0.5	0.5I	0.6	0.6I	0.7	0.7I
Random	19.77	.99	.82	1.35	1.75	.64	.62
Triangular	19.72	1.03	.85	1.43	.75	.46	.46
Regular	18.77	1.15	.91	1.58	.76	.54	.53
Aggregate	19.90	1.47	1.18	1.95	.96	.73	.72
Double clumped	21.84	1.08	.89	1.20	.86	.88	.88
[illegible]	21.53	1.20	.93	1.59	.77	.69	.67

[1] "I" indicates the use of interpolation.

Discussion

In general, the small sample size situation is where PDE's are most efficient for estimating density (see, for example, Pollard 1971 and Holgate 1964) if the underlying assumptions are met. However, small sample sizes are not very effective for estimating percentiles. This problem with using the PBK estimator for small sample sizes is recognized by Patil et al. (1979) and Bythe (1982). Interpolation is one means to reconcile the small sample size inefficiencies of the PBK estimator with the general efficiency of PDE's at small sample sizes. The simulation results presented in the previous section indicate that interpolation consistently improves the estimation properties of the PBK estimator across a variety of spatial patterns where the assumptions for many PDE's would not be met.

Acknowledgment

The authors thank Ken Burnham for his helpful review and comments.

Literature Cited

Bythe, K. 1982. On robust distance-based intensity estimators. Biometrics 38: 127–135.

Diggle, P. J. 1975. Robust density estimation using distance methods. Biometrika 62: 39–48.

Holgate, P. 1964. The efficiency of nearest neighbour estimators. Biometrics 20: 647–649.

Loftsgaarden, D. O.; Quesenberry, C. P. 1965. A nonparametric estimate of multivariate density function. Annals of Mathematical Statistics 36: 1049–1051.

Patil, S. A.; Burnham, K. P.; Kovner, J. L. 1979. Nonparametric estimation of plant density by the distance method. Biometrics 35: 597–604.

Patil, S. A.; Kovner, J. L.; Burnham, K. P. 1982. Optimum nonparametric estimation of population density based on ordered distances. Biometrics 38: 243–248.

Pollard, J. H. 1971. On distance estimators of density in randomly distributed forests. Biometrics 27: 991–1002.